What Does the System Usability Scale (SUS) Measure?

Validation Using Think Aloud Verbalization and Behavioral Metrics

Mandy R. Drew(✉), Brooke Falcone, and Wendy L. Baccus

WeddingWire Inc., Chevy Chase, MD 20815, USA
mdrew@weddingwire.com

Abstract. The System Usability Scale (SUS) is widely used as a quick method for measuring usability; however, past research showed there is only a weak relationship between SUS scores and one behavioral usability measure, and alternatively, SUS corresponds more strongly with user preference. This suggests that the underlying constructs of the SUS may not be well understood. In this study, participants were asked to think aloud while completing a usability test and filling out the SUS. Correlations showed no relationship between behavioral performance and SUS scores. Instead, a relationship was observed between SUS scores and perceived success. Furthermore, participants described a variety of reasons for selecting their SUS responses that were unrelated to the usability of the system, which we have termed *rationalizations*. This suggests that the SUS is constructed of a combination of experiential components, including attitudinal perceptions. Consequently, SUS scores may be more helpful as a tool for comparison (between competitors, iterations, etc.,) or when used in conjunction with formative usability testing methods to provide a holistic view of real and perceived user experience.

Keywords: System Usability Scale · SUS · Think-aloud protocol
Usability testing · User experience

1 Introduction

1.1 Usability Testing

Rooted in ancient principles of ergonomic design, (Marmaras et al. 1999) usability describes how effective, efficient, and satisfying a product is when used to complete the tasks for which it was designed (International Organization for Standardization 1998). While a wide range of attitudinal research methods can be used to understand users' stated beliefs, a large body of evidence shows self-reported claims and speculation are unreliable assessments of usability (Frokjaer et al. 2000; Nielsen and Levy 1994). Asking users to evaluate a product based on whether or not they like it is ineffective because users sometimes prefer designs that actually prevent them from achieving their goals (Andre and Wickens 1995), and requiring users to evaluate an experience

A. Marcus and W. Wang (Eds.): DUXU 2018, LNCS 10918, pp. 356–366, 2018.
https://doi.org/10.1007/978-3-319-91797-9_25

retrospectively is unreliable because users tend to remember and make judgments based on the most recent or intense parts of the experience they encountered (Kahneman et al. 1993). Thus, attitudinal research methods can be effective in measuring user perception but cannot accurately determine usability or pinpoint areas for improvement (Nielsen 1993).

Instead, it is widely agreed that watching participants interact with a product while attempting to use it is a much better measure of overall usability and can also diagnose specific usability problems that exist (Nielsen and Levy 1994). In order to discover users' cognitive processes as they are completing tasks, it is helpful to record them speaking their thoughts aloud during test administration (Ericsson and Simon 1980; Nielsen 1993). Such think-aloud protocol usability studies have proven valuable in identifying design flaws and comprehension gaps users may lack the ability to independently articulate, shedding light on opportunities for improvement (Nielsen and Levy, 1994) without requiring recall on the part of the user.

The implementation of traditional usability studies on digital products has evolved from what was once a fairly complex, expensive process. In the past, studies were conducted in-person with 30–50 participants (Barnum 2002) within the confines of a usability lab outfitted with sophisticated video recording equipment (Nielsen 1993). Modern research reveals testing only 5–7 representative users uncovers about 80% of problems users face, (Nielsen and Landauer 1993) and technological advances have made it possible to reach and test those individuals remotely without the need for a recruiter, moderator, or expensive hardware, decreasing the time and expense it takes to watch target users interact with a system (Bergel et al. 2002). Despite this advance in testing efficiency, increased adoption of the agile software development methodology, which advocates for fast and continuous product delivery, (Beck et al. 2001) may tempt some researchers to forgo even the more streamlined version of remote, unmoderated usability testing available today in favor of self-reported measures, such as the SUS, for speed.

1.2 The System Usability Scale

Over three decades ago, the SUS was developed as a subjective usability measure capable of being administered quickly after users worked through evaluation tasks (Brooke 1996). The questionnaire asks users to rate their level of agreement with statements covering a variety of usability characteristics such as the system's complexity and any support or training participants believe is required to use it effectively. Brooke asserts the simple 10-item post-test questionnaire can quickly assess a product's usability without the need for complicated analysis.

Advantages. The SUS has many benefits that make it a popular choice for usability assessment. In addition to being fast, free, and easy to administer, it is considered fairly simple for participants to answer. This is an important consideration, given participants may have struggled for some time to complete frustrating tasks during the testing process. To account for bias due to fatigue or inattention, the questions alternate between positive and negative, providing a single score which, when translated to a familiar "university grade," is commonly understood and easily shared among project

stakeholders (Bangor et al. 2008). Finally, the SUS can be applied to a wide range of technologies and has been used to assess the usability of hardware, software, websites, and mobile devices (Brooke 2013).

Disadvantages. The advantages of SUS coupled with the fast pace of modern product development cycles has made it an industry standard with references in over 1,300 articles and publications (Usability.gov, n.d.), but relying on SUS as a sole measure of usability is problematic for a number of reasons.

First, more recent evidence suggests that the SUS primarily measures subjective user perception. For instance, one study shows that users who give a system a high SUS score also tend to give that same system a high Net Promoter Score (NPS), which indicates that they are very likely to recommend the system to a friend or family member (Sauro 2011). While this means the SUS could be a good indicator of preference, that does not necessarily relate to product usability because users do not always like usable designs (Andre and Wickens 1995). Furthermore, only a very small relationship has been shown between SUS results and the ability to effectively complete usability tasks without making any errors (Peres et al. 2013).

Several studies also highlight how users and researchers can misinterpret items on the SUS. For example, in approximately 13% of questionnaires, the alternating positive and negative items are responsible for incorrect responses and coding errors. (Sauro and Lewis 2011). In addition, the use of the word "cumbersome" in item 8 has been shown to cause confusion for about 10% of all SUS respondents (Bangor et al. 2008; Finstad 2006).

Another notable limitation is that, because SUS questionnaires are administered after the initial test is complete, users may encounter the peak-end effect and evaluate the system based on the most recent or intense parts of their experience (Kahneman et al. 1993). This is problematic because users miss reporting specific events that could be pertinent to the system's overall usability and can result in missed opportunities to pinpoint potential areas for design improvement.

Finally, the assumption that users provide answers to SUS questions based exclusively on the experience they are meant to evaluate has never been validated. In fact, it has been shown that prior experience with a system can lead to more favorable SUS scores (McLellan et al. 2012), which suggests that outside influences can affect how users respond to the questionnaire.

All these factors could potentially explain the negative skew described in a 2008 study which found that the mean scores from SUS questionnaires in over 200 tests were more favorable than the actual test success rates (Bangor et al. 2008).

1.3 Present Study

Previous research suggests that the SUS may better reflect users' perceived success, rather than a system's overall usability (Peres et al. 2013). On the surface, SUS questions appear to cover a variety of aspects essential to usability, such as complexity, learnability, and likelihood of repeat use (Brooke 1996). Thus, the majority of SUS respondents likely recognize the purpose of the survey is to elicit an accurate understanding of their experience with the system in question. However, such "face validity" only concerns judgments made about the survey after it was created but does not

necessarily indicate the survey was constructed in such a way as to measure what it asserts (Nunnally 1978). The primary goal of this study is to explore SUS' construct validity; that is, to discover how users justify the SUS responses they choose in order to better understand the disconnect between users' behavior and their perceived success, as reflected by their SUS ratings.

Hypothesis. We expected that participants who exhibited poor behavioral usability measures on a traditional think-aloud protocol usability test but submitted high SUS scores for the same experience would provide verbal rationalizations to explain the SUS answers they chose. Exploring what participants say could provide insights into their cognitive processes in order to better understand how SUS responses relate to task success.

2 Methods

2.1 Participants

Twenty participants from the United States, (17 females, 3 males) with a mean age of 27.1 (SD = 4.06) were recruited from an online panel. All had prior experience with think-aloud protocol and were familiar with the testing platform. Participants were screened for domain-related qualities to ensure they represented real users of the websites in question: (1) engaged to be married or recently married, (2) had never used either of the websites to be evaluated, and (3) had created a wedding registry online. Participants earned a $10 incentive.

2.2 Materials

The study was unmoderated and conducted remotely via Usertesting.com, an online platform that streamlines and automates the recruitment, implementation, and analysis of remote, unmoderated usability studies.

2.3 Procedures

Panelists received invitations to complete a screener survey to determine adherence to the inclusion criteria. The first 20 participants to meet the criteria were prompted to launch the test from their computers and used their system's hardware to record their screen actions and voices as they followed the test tasks.

All twenty participants attempted to complete two tasks on two different websites and the presentation of each website was counterbalanced across participants. In order to measure participants' perception of success, they were asked after each task if they felt they completed it successfully. After testing each website, the SUS questionnaire was administered to evaluate their satisfaction with each system.

The Usertesting.com platform displayed written test instructions and tasks on-screen, and participants' voices and screen actions were recorded throughout test administration. The perceived effectiveness question displayed on-screen after the completion of each task and the SUS questions appeared on-screen after each website

was evaluated. Participants selected the appropriate response to the perceived effectiveness and SUS questions from a series of radio buttons. Upon completion of the test, each participant's video, demographic information, and responses to perceived effectiveness and SUS questions were uploaded to the Usertesting.com platform where they were accessed and analyzed by researchers.

2.4 Design

A within-subjects design was used. To mitigate the effects of learning bias and ordering effects, the test was counterbalanced to randomize the order of websites evaluated.

2.5 Measures

In order to gain a holistic understanding of the usability of both experiences, we measured Effectiveness, Efficiency, and Satisfaction (International Organization for Standardization 1998). To understand the reasons why users chose the SUS scores they selected, we created a new metric we called Rationalization.

Behavioral. A previous study calculated the Effectiveness behavioral metric as the percent of test tasks participants completed without making any errors. (Peres et al. 2013). Our study followed the more common practice of assigning a binary value as the Task Success rate (Sauro and Lewis 2005), and also by calculating the total number of Unique Errors each participant made (Sauro and Lewis 2016). This method allowed us to diagnose specific UI problems and also better understand participants' frustration levels with two experiences that were objectively difficult to use. Efficiency (or Time on Task) was calculated by the number of seconds participants spent on each task.

Attitudinal. Satisfaction, or comfort using the system, was determined by tabulating each response submitted to the SUS questionnaire and scoring it according to the methods outlined by John Brooke (Brooke 1996). We measured participants' Perceived Success by asking them whether or not they felt they completed each task successfully and assigning a binary value to their responses. Additionally, we introduced a new metric called Rationalization, in which we observed participants verbally rationalize their decisions for selecting each SUS response.

In order to quantitatively assess the correspondence between user's subjective verbalizations and their SUS ratings, we applied a procedure inspired by the eight-step process for coding verbal data developed by Chi (1997). One researcher independently organized the rationalizations into categories (see Table 2), and another researcher assigned a binary value to indicate in which of the six resulting categories each rationalization best applied: (1) No Rationalization, (2) Blames Self, (3) Minimizes Issues, (4) Good by Comparison, (5) No Basis for Comparison, and (6) Halo Effect. Interrater reliability could not be calculated due to the use of only one rater.

In the event that a participant failed to elaborate on their thought process, the "No Rationalization" category was included. Some participants blamed poor performance on their own unfamiliarity with the subject matter or lack of technological ability, whose responses comprise the "Blames Self" category. This example of social desirability bias stems from the need for social approval and is a common limitation for

many user research activities (Nancarrow and Brace 2000). The "Minimizes Issues" category consists of those participants who minimized significant usability issues that caused errors because they considered the overall experience to be positive. The "Good by Comparison" category refers to participants who evaluated the second website in comparison to the first they encountered. The "No Basis for Comparison" category was created to exemplify users that could exhibit a lack of basis for comparison by interpreting an objectively poor experience, such as spending ten or more minutes attempting to complete a task, as adequate. While this rationalization was not observed by the researcher who rated the verbalizations for this study, it should be considered as representative of future samples. Lastly, the halo effect (Berscheid and Walster 1974; Angeli and Hartmann 2006) was observed when participants' valued aesthetics over practicality.

3 Results

3.1 Usability Measures

Correlations for the behavioral and attitudinal usability measures are reported in Table 1. The results indicate that the strongest correlation exists between Time on Task and Errors, such that participants who experienced more errors took longer to complete tasks. Also, there was a strong positive relationship between Task Success and Perceived Success, such that those who experienced more Task Success also scored highly in Perceived Success. Lastly, a weaker, but still moderate relationship was seen between SUS Scores and Perceived Success, indicating that SUS scores increase as perceived success increases.

Table 1. Correlation matrix of behavioral and attitudinal usability measures (N = 40).

	SUS	Errors	Task success	Perc. suc.	ToT
SUS	–				
Errors	.03	–			
Task success[a]	.29	−.07	–		
Perc. suc.[b]	.51**	−.28	.66**	–	
ToT	−.09	.68**	.19	−.13	–

Note: ** $p < .01$, * $p < .05$, n = 40.

3.2 Rationalizations

A Mann-Whitney test indicated that there was not a significant difference in Task Success between those who rationalized and those who did not ($U = 178.00$, p = ns). A Mann-Whitney test indicated that there was not a significant difference in Perceived Success between those who rationalized and those who did not ($U = 182.00$, p = ns).

A MANOVA indicated there was a statistically significant difference in Errors (see Fig. 1) based on whether a user rationalized $F(1, 40) = 4.522$, $p < .005$. Participants who rationalized committed more errors than participants who did not rationalize.

Table 2. Rationalizations

	Frequency count	% of total	Examples
None	325	.81	
Self-blame	13	.03	"I get confused easily" "My lack of knowledge of wedding registries… makes me inexperienced" "I guess I was doing tasks too fast"
Minimization	10	.10	"Once you learn what you are doing it gets easier. It's just a matter of getting over the first hump which is ridiculous and annoying" "Maybe I could've gotten it with more playing with it" "There were [sic] some weird programming with the buttons, but other than that it was fantastic" "It seemed easy, but I couldn't figure out how to add the registry"
Good by comparison	14	.04	"Easier than [the other website]"
No basis for comparison	0	.00	
Halo effect	10	.03	"Everything matches the beautiful blue color" "I would use the other functions"

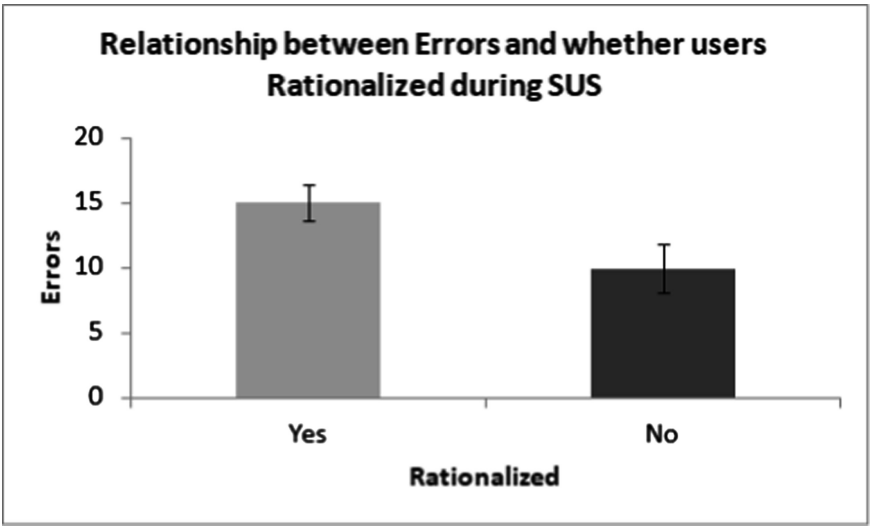

Fig. 1. Average error frequency committed by participants who rationalized and those who did not. Error bars indicate standard error for each group.

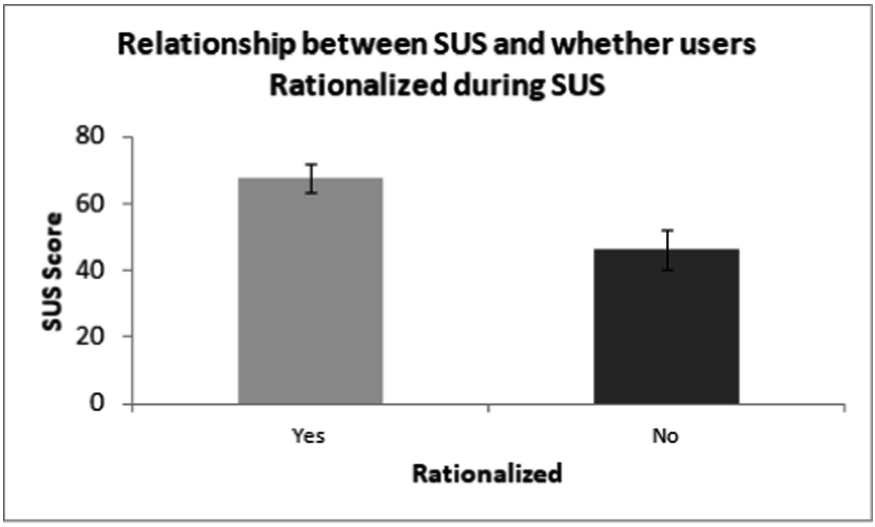

Fig. 2. Average SUS scores reported by participants who rationalized and those who did not. Error bars indicate standard error for each group.

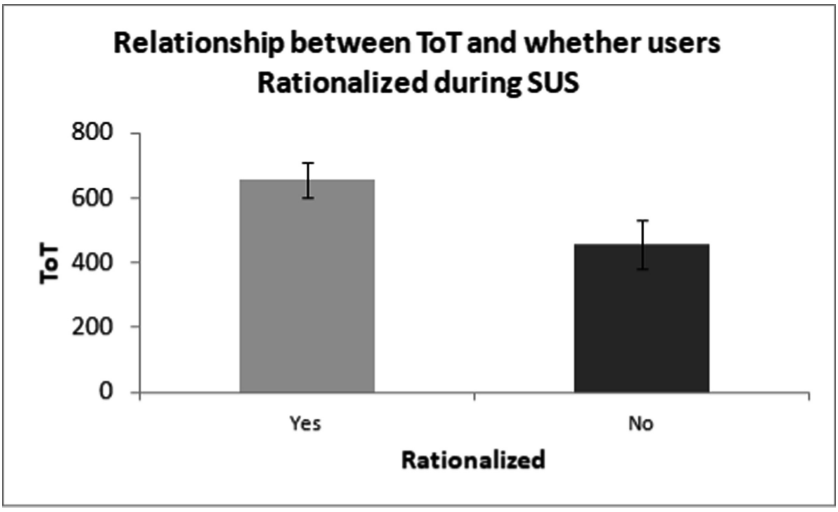

Fig. 3. Average Time on Task of participants who rationalized and those who did not. Error bars indicate standard error for each group.

In addition, there was a statistically significant difference in SUS scores (see Fig. 2) based on whether a user rationalized $F\ (1,\ 40) = 8.699$, $p < .005$, where participants who rationalized reported higher SUS scores than those who did not rationalize. There was also a statistically significant difference in ToT based on whether a user rationalized $F\ (1,\ 40) = 4.81$, $p < .005$ indicating that those who rationalized spent more time completing tasks than those who did not rationalize (Fig. 3).

4 Discussion

4.1 Interpretation

These findings support previous research that suggests SUS is a measure of user perception, and not actual usability. Regardless of their performance, participants who believed they completed their tasks successfully gave the system a higher SUS score. Participants who struggled to complete their tasks efficiently and effectively often rationalized their performance for reasons unrelated to system usability and gave the system a higher SUS score than an objective review of their experience warranted.

Comparison to Previous Study. The small relationship previously discovered between SUS scores and effectiveness (Peres et al. 2013) is not replicated in these results. That study compared the percent of tasks users completed without any errors to their corresponding SUS scores, and suggested the resulting correlation provided evidence that the SUS could be used to ordinally compare the usability of multiple systems (Peres et al. 2013). That "error-free" method for calculating effectiveness was impractical for this study, given the high level of difficulty users encountered; in fact, none of the tasks for this study were completed without errors. Thus, we determined effectiveness using three different measures: (1) Task Success, (2) Perceived Success, and (3) Unique Errors. No relationship was discovered between Task Success and SUS, but the positive association between SUS scores and Perceived Success suggests that SUS reflects user perception. Participants who believed they completed their tasks successfully gave the system a higher SUS score, regardless of their actual performance. This could explain the relationship described in Peres et al. (2013). Users understood they completed tasks without making mistakes, and thus, chose more favorable SUS responses. Despite the fact that a large proportion of participants did not rationalize while performing the SUS, it is possible that they followed a similar thought process, without verbalizing. Of those who did rationalize, our results demonstrate that SUS scores do not relate to task success; therefore, we recommend it be used in conjunction with standard usability metrics to comprehensively explain the user's experience with a system.

4.2 Practical Applications

The rationalizations observed in this study characterize the type of noise that can cloud objective usability measurement using the SUS and could contribute to the negative skew observed in Bangor et al. (2008). Attempting to gain insight into why users select the SUS responses they choose for each individual study could limit the adverse effects caused by this noise but would diminish the advantages of using SUS as a "quick and dirty" method of usability assessment. Leveraging both formative and summative measures is a commonly recommended best practice (Nielsen and Levy 1994) to gain a holistic understanding of a product's performance, and formative assessments, such as think-aloud protocol usability tests, could mitigate the effects of cognitive bias because researchers can watch what users do with a system while simultaneously discounting what they say about it. However, the effects of employing methods simultaneously on the outcome variables and, ultimately, the conclusions drawn from such research has not been well explored and may introduce interaction effects that could limit the ability to interpret and generalize the results.

A more practical application of SUS scores could be as an instrument of the test-retest reliability method (Carmines and Zeller 1979). The SUS questionnaire could be administered after initial traditional usability testing on the first iteration of a product and the resulting score could be used as a benchmark by which subsequent design iterations could be evaluated. In the event that future product versions garner lower SUS scores, traditional usability testing can again be employed to find out why. This addresses the need for quick product evaluation and continuous delivery while also maintaining research integrity by adhering to industry standard best practices.

4.3 Future Research

It is important to note that the act of recording users during SUS administration could have affected their responses (Macefield 2007), and may cause users to overthink, which John Brooke cautioned against (Brooke 1996). This is beyond the scope of this paper but is an appropriate area for follow-up study.

4.4 Limitations

The present results are exploratory findings uncovered while conducting a larger study aimed to answer an alternative research question. Future studies should explicitly design and test the conclusions incidentally described here.

Future studies employing the Rationalization coding scheme constructed in this study should seek to validate the reliability of the measure between raters using inter-rater reliability (IRR) (Hallgren 2012).

References

Andre, A., Wickens, C.: When users want what's not best for them. Ergon. Des. Q. Hum. Factors Appl. **3**(4), 10–14 (1995)

Angeli, A., Hartmann, J.: Interaction, usability, and aesthetics. In: Proceedings of the 6th Annual ACM Conference on Designing Interactive Systems, pp. 271–280. ACM, New York (2006)

Beck, K., Beedle, M., Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R., Mellor, S., Schwaber, K., Sutherland, J., Thomas, D.: Manifesto for Agile Software Development (2001). http://www.agilemanifesto.org/. Accessed 29 Jan 2018

Bangor, A., Kortum, P.T., Miller, J.T.: An empirical evaluation of the system usability scale. Int. J. Hum.-Comput. Interact. **24**(6), 574–594 (2008)

Barnum, C.M.: Usability Testing and Research. Longman, New York (2002)

Bergel, M., Fleischman, S., McNulty, M., Tullis, T.: An empirical comparison of lab and remote usability testing of web sites. In: Usability Professional Association Conference (2002)

Berscheid, E., Walster, E.: Physical attractiveness. In: Berkowitz, L. (ed.) Advances in Experimental Social Psychology, vol. 7. Academic Press, New York (1974)

Brooke, J.: SUS: a "quick and dirty" usability scale. In: Usability Evaluation in Industry, pp. 189–194. Taylor & Francis, London (1996)

Brooke, J.: SUS: a retrospective. J. Usability Stud. **8**(2), 29–40 (2013)

Carmines, E., Zeller, R.: Reliability and Validity Assessment. SAGE Publications, Thousand Oaks (1979)
Chi, M.: Quantifying qualitative analyses of verbal data: a practical guide. J. Learn. Sci. **6**(3), 271–315 (1997)
Ericsson, K., Simon, H.: Verbal reports as data. Psychol. Rev. **87**(3), 215–251 (1980)
Finstad, K.: The system usability scale and non-native English speakers. J. Usability Stud. **4**(1), 185–188 (2006)
Frokjaer, E., Hertzum, M., Hornbaek, K.: Measuring usability: are effectiveness, efficiency, and satisfaction really correlated? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 345–352. ACM Press, New York (2000)
Hallgren, K.: Computing inter-rater reliability for observational data: an overview tutorial. Tutor. Qual. Methods Psychol. **8**(1), 23–34 (2012)
International Organization for Standardization: ergonomic requirements for office work with visual display terminals (VDTs) – Part 11: guidance on usability (ISO Standard No. 9241-1.) (1998). https://www.iso.org/standard/16883.html. Accessed 29 Jan 2018
Kahneman, D., Fredrickson, B., Schreiber, C., Redelmeier, D.: When more pain is preferred to less: adding a better end. Psychol. Sci. **4**(6), 401–405 (1993)
Macefield, R.: Usability studies and the Hawthorne effect. J. Usability Stud. **2**(3), 145–154 (2007)
Marmaras, N., Poulakakis, G., Papakostopoulos, V.: Ergonomic design in ancient Greece. Appl. Ergon. **30**(4), 361–368 (1999)
McLellan, S., Muddimer, A., Peres, S.: The effect of experience on system usability scale ratings. J. Usability Stud. **7**(2), 56–67 (2012)
Nancarrow, C., Brace, I.: Saying the right thing: coping with social desirability bias in marketing research. Bristol Bus. Sch. Teach. Res. Rev. **3**(11), 1–11 (2000)
Nielsen, J., Landauer, T.: A mathematical model of the finding of usability problems. In: Proceedings of ACM INTERCHI 1993 Conference, Amsterdam, The Netherlands, pp. 206–213 (1993)
Nielsen, J.: Usability Engineering, 1st edn. Academic Press Inc., Cambridge (1993)
Nielsen, J., Levy, J.: Measuring usability - preference vs. performance. Commun. ACM **37**(4), 66–75 (1994)
Nunnally, J.: Psychometric Theory, 2nd edn. McGraw-Hill, New York (1978)
Peres, S., Pham, T., Phillips, R.: Validation of the system usability scale (SUS): SUS in the wild. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 57, no. 1, pp. 192–196 (2013)
Sauro, J.: A Practical Guide to the System Usability Scale: Background, Benchmarks & Best Practices. Measuring Usability LLC, Denver (2011)
Sauro, J., Lewis, J.: Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 49, no. 24, pp. 2100–2103 (2005)
Sauro, J., Lewis, J.: Quantifying the User Experience: Practical Statistics for User Research, 2nd edn. Morgan Kaufmann, Cambridge (2016)
Sauro, J., Lewis, J.: When designing usability questionnaires, does it hurt to be positive? In: Proceedings of ACM SIGCHI, pp. 2215–2223. ACM, New York (2011)
Usability.gov: System Usability Scale (SUS). http://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html. Accessed 31 Jan 2018